DOCUMENT RESUME

ED 281 858 TM 870 237

AUTHOR Wijnstra, Johan M.; Eggen, Theo J. H. M.
TITLE Design of the Dutch National Assessment Program in Education.
PUB DATE Apr 87
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Curriculum Evaluation; Educational Assessment; Elementary Education; *Elementary School Curriculum; Foreign Countries; Item Analysis; *Latent Trait Theory; *Mathematics Tests; National Programs; Questionnaires; Research Design; Sampling; Standardized Tests; *Test Construction; *Testing Programs; Test Items
IDENTIFIERS Calibration; *Netherlands; Rasch Model; Second International Mathematics Study

ABSTRACT

The operational preparations for the Dutch national assessment program in education started in 1986. The program's aim is to periodically describe the contents of the curriculum in use and the attainment of 8-year-old and 11-year-old students on a national basis by means of surveys. In this paper some of the guiding principles in instrument construction, sampling, and data collection are described. The instruments are of two types: (1) questionnaires to make an inventory of the topics actually taught and the variation in the intensity of teaching; and (2) tests to measure the performance of 8-year-old and 11-year-old students in narrowly defined content domains. Data collection will start in 1987 in the field of mathematics. The design for this assessment is described in more detail with special attention to a problem in the application of item response theory in scale construction: the invariance of the location parameters in relation to the opportunity to learn the prerequisite concepts and skills. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281858

DUTCH NATIONAL ASSESSMENT PROGRAM IN EDUCATION PPON

Address: National Institute for Educational Measurement

P.O. Box 1034
6801 MG Arnhem
The Netherlands

Design of the Dutch national assessment program in education

Johan M. Wijnstra & Theo J.H.M. Eggen

National Institute for Educational Measurement
Arnhem, the Netherlands

Abstract

The operational preparations for the Dutch national assessment program in education started in 1986. The program's aim is to periodically describe the contents of the curriculum in use and the attainment of eight and eleven year old students on a national basis by means of surveys. In this paper some of the guiding principles in instrument construction, sampling, and data collection are described. Data collection will start in 1987 in the field of mathematics. The design for this assessment is described in more detail with special attention to a problem in the application of item response theory in scale construction: the invariance of the location parameters in relation to the opportunity to learn the prerequisite concepts and skills.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. M. Wijnstra

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper prepared for the Annual Meeting of the American Educational Research Association, April 20 - 24, 1987, Washington, DC

BEST COPY AVAILABLE

TM 870 237

PPON – the Dutch national assessment program in education, commissioned by the Minister of Education and Science, is carried out as a joint project of the Institute for Educational Research in the Netherlands and the National Institute for Educational Measurement

Introduction

After several years of discussion at the national policy level and the conduct of a pilot study in primary education, last year the Dutch national assessment program in education was launched. The Minister of Education and Science commissioned the National Institute for Educational Research in the Netherlands and the National Institute for Educational Measurement to carry out the assessment program as a joint project. Until further notice the project is restricted to an assessment of the primary education system.

The program's aim is to periodically describe the contents of the curriculum in use and the attainment of eight and eleven year old students on a national basis by means of surveys, so as to provide a basis for a more rational discussion about the contents and the level of primary education. In a period of four years the whole curriculum should be covered, including social studies, science, physical education, art, and music.

In this paper some of the guiding principles in the design of the assessment program will be described. Data collection will start this spring term with the assessment of mathematics at the end of primary education. The design for this assessment is described in more detail, with special attention to a problem in the application of item response theory in scale construction: the invariance of the location parameters in relation to the opportunity to learn the prerequisite concepts and skills.

Instruments

Instrument development starts with the construction of a framework, describing the structure of the learning area under study. The frameworks are primarily content oriented and rather detailed, so as to

provide the basis for the description of the curriculum in use and the performance of students in meaningful units for educators and curriculum specialists, more or less comparable to what Bock, Mislevy, and Woodson (1982) call 'indivisible curricular elements'. The frameworks, which provide a rationale for the instrument construction, are presented for comment to several groups of relevant respondents, like teachers and curriculum specialists, of course, but also representatives of societal organizations such as political parties, unions, and employers' associations. The frameworks do not imply that all topics included are common practice to the same degree in Dutch schools. Such a restriction beforehand might curtail a discussion of the results with respect to desirable developments in the contents of education too much.

The frameworks provide the basis for the construction of two types of partially complementary instruments:

- questionnaires to make an inventory of the topics actually taught and the variation in the intensity of teaching;
- tests to measure the performance of students at two age levels (eight and eleven year olds) on narrowly defined content domains.

The inventory of the contents taught constitute data in their own right. In addition, these data can be used to explore relations between performance and content of the curriculum taught.

The tests are meant to cover a wide range of skills and concepts. In the language assessment, for instance, reading and writing proficiency as well as speaking and listening skills will be assessed. Each individual test will cover a rather small content domain. The scope of the domains is comparable to the curricular units used in the California Assessment Program (cf. Pandey & Carlson, 1983). The test items and exercises will be developed by committees of experienced

teachers under the direction of project staff members. A high degree of curricular validity is aspired to in the test construction. If the curricular validity of multiple choice items is questionable, open-ended items and exercises will be used. In the mathematics assessment, for example, more than 95% of the items will use the open-ended format.

As item response theory has several advantages in large-scale assessment (cf. Bock & Mislevy, 1981; Bock et al., 1982; Messick, Beaton, & Lord, 1983), these models will be applied in scale construction, whenever dichotomously scored items are used. The application of the one or two-parameter logistic model is planned. The rather detailed frameworks have several advantages in this context, e.g. changes in curriculum can easily be catered for without losing comparability over time on unaffected scales, while the probability of arriving at a reasonable model fit is greatest (cf. Traub & Wolfe, 1981; Hambleton & Swaminathan, 1985).

Populations and sampling

Compulsory education starts at age five in the Netherlands, but almost all children enter primary school on a voluntary basis at the age of four. The primary school contains eight grades, the first two grades more or less comparable to a Kindergarten program. The Dutch national assessment program will collect data about the curriculum content and achievement of students about midway and at the end of primary school. Since grade alone is not a very good basis to categorize students, the student populations are defined as the students in a given grade (5 or 8) and all students who could have been in that grade according to their age (students of 8 and 11 years old, respectively, on September 30 of the schoolyear in which the data are being collected). This

provides the opportunity to define three relevant populations:

- the students in a given grade, irrespective of age;
- the students of the modal age in a given grade;
- the students of a given age, irrespective of grade.

The population of schools consists of all primary schools in the Netherlands. Although about two thirds of the schools have non-public governing bodies, financially almost all schools are completely publicly maintained. To determine the number of teachers a school is entitled to hire, a weighted number of students is calculated. The weight for lower-class indigenous Dutch students is 1.25, for example, and 1.90 for minority students.

The ratio of the weighted number of students and the nominal number is used to stratify the population of schools into three groups:

* $\leq$ 1.05 : 47% of the schools;
* 1.06 - 1.15: 35% of the schools;
* $>$ 1.15 : 18% of the schools.

In each stratum the samples will contain the same number of schools, in principle, in such a way that the mean test scores in each stratum can be estimated with a precision of 0.075. In order to retain representativity on the stratification variable, systematic random samples are planned, with two replacements for each sampled school.

All students available in the sampled schools participate in the testing sessions conducted by specially trained project assistants. To reduce the clustering effect, each student takes only a few of the tests, at the same time rotating the tests over students within schools. In principle, testing takes no longer than half a school day. Therefore, the sample size will also be dependent on the number of tests to be used and the time these tests take.

Mathematics assessment at the end of primary education

The framework for the mathematics assessment at the end of primary school contains 27 topics of rather narrowly defined curriculum components like:

- fractions: basic concepts;
- addition and subtraction of fractions and rational numbers;
- multiplication and division of fractions and rational numbers;
- fractions: applications.

In the curriculum inventory each topic is illustrated with three items from the corresponding tests, varying in difficulty. Aiming at the average student in grade 8, the following questions are asked to the classroom teacher about each topic:

- Have the concepts and skills necessary to solve each of the exemplar items been taught to the average student in your class?
- How much time did you spend in teaching this topic this schoolyear?

In order to estimate the time spent, first the total time spent in mathematics is to be estimated. The time spent per topic is estimated with the help of 100 cardboard chips, each chip representing 1% of the total time spent to mathematics. Complementary to the teacher questionnaire about the average student, on the list with the names of the students the question is asked whether each of the students has been taught the concepts and skills to solve the easiest exemplar item of each topic.

For each topic a scale is to be constructed, using item response theory. With one exception, between 16 and 20 items have been constructed for each topic. The first test, on basic mathematics skills, contains 32 items, which are administered in a structurally incomplete design. Each student takes only two blocks of eight items of this topic. The blocks are systematically rotated over students. All

other topics are also administered in a structurally incomplete design, but in these cases a priori collateral information is used to assign tests to students, so as to maximize the information and to minimize the embarrassment to students. The items on each topic are divided into two test versions, a relatively easy one (A) and a relatively hard one (B), with an overlap of eight items. Whether a student receives a booklet containing A or B tests, is decided by the classroom teacher. All students who are one or more grades behind, compared with their age mates, take the easy test versions, of course.

In addition to the basic mathematics skills test, each student takes items on four or five topics. Basically, six A and six B booklets are needed. So as to reduce order effects, each test has been produced in two item orders, while at the same time the booklets have been assembled in two different topic orders. Consequently, 48 booklets have been produced: 24 A and 24 B booklets.

The tests will be administered in a sample of 350 schools during the month of May this year. The total sample consists of systematic random samples of 100 schools from each of the first two strata and 150 from the third. The latter sample size has been increased in order to simultaneously collect data for an evaluation program on educational priority areas. The schools in the sample count about 8000 students who are eligible for the assessment program. Each item will be taken by at least 500 students, while the overlapping items in the A and B tests will be taken minimally by double that number. The sample size is sufficient to estimate the mean test scores in each stratum with a precision of 0.075, even when the intraclass correlation exceeds a value of 0.20.

In the calibration of the items and the estimation of the (distribution of) ability parameters a special problem with respect to

model fit might arise from the circumstances that students participate who are one or two grades behind. Although these students take the easy test versions, this could mean that the concepts and skills necessary to solve some problems have not yet been taught. This might pose problems for the invariance of the location parameters of the items. That is the reason why each student is judged on the prerequisite knowledge to solve one of the easiest items on the scale. In the event that the inclusion of these students distorts the stability of the item parameters, these students can be excluded from the analyses and their proportion in the population can be estimated. This problem will be elaborated upon in the final section.

Opportunity to learn and invariance of location parameters

Wilmot and Tuson (1983) investigated this problem with respect to mathematics in the context of the British Assessment of Performance Unit (APU), using the Rasch model. They linked test performance data to teacher questionnaire data about exposure of their students to the topics tested and compared the relative item difficulties in the total sample and the subset of exposed students. They concluded: 'In the calibration of items the inclusion of pupils judged by their teachers not to be exposed to some of the material in the tests taken makes very little difference to the relative item difficulties of most items.' (p. 100). However, on the test with the lowest exposure index (modern geometry) they report: 'Although there is general agreement between the two sets of relative item difficulties, in some cases there is a complete reversal of rank ordering (p. 96). So, some care is in order.

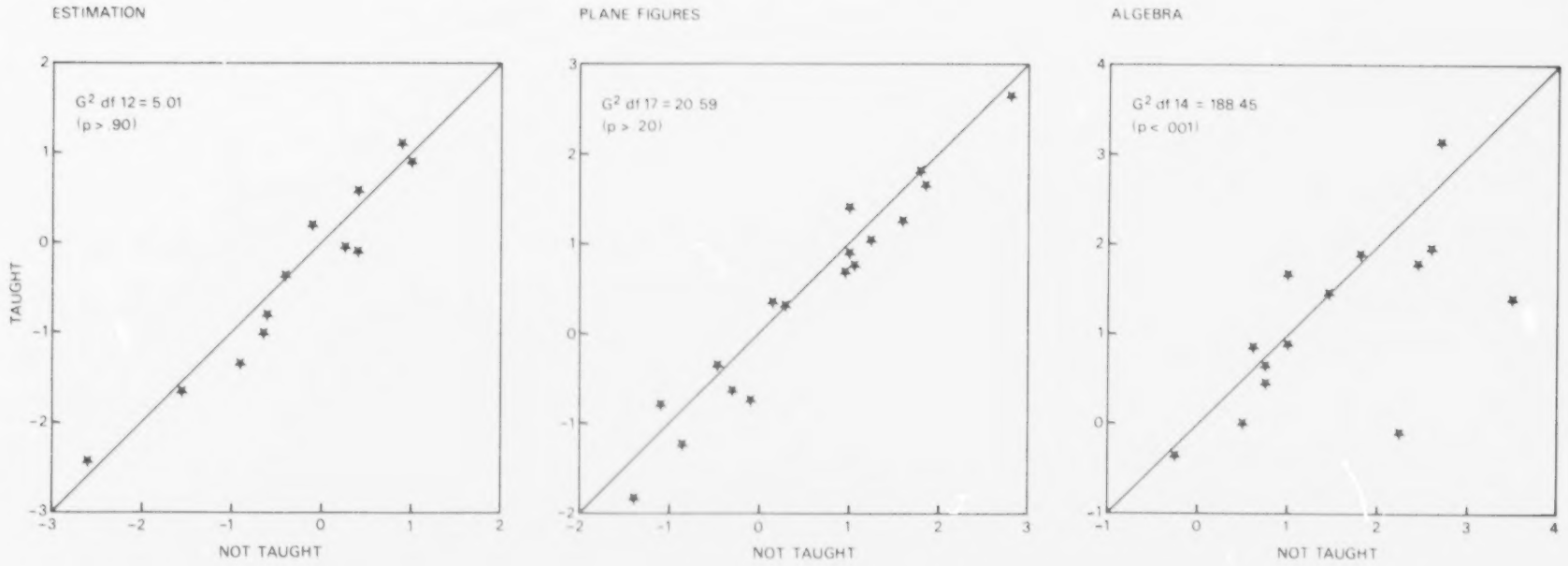
To further explore the problem, we decided to analyse some subtests from the Second International Mathematics Study (13-year-olds) of the International Association for the Evaluation of Educational Achievement

(IEA). The IEA data were chosen because data on opportunity to learn are available. In some cases students were also asked these questions, but we only used data from teacher questionnaires. Three subtests were included in the analyses: Estimation, Algebra and Plane Figures. These are of about the same length as the tests we are going to use in our mathematics assessment and vary in the degree of curriculum dependency. For several reasons samples of about 1000 students from the Swedish data set were selected.

The IEA tests were administered in a structurally incomplete design. Some items were taken by all students, others by about half. To study the influence of opportunity to learn on the location parameters, two sets of pseudo-items were created by splitting the data set into two sets, one containing the items taken by students who had been exposed to the topic during the (then) current school year or earlier, the other with the same items taken by students who had had no opportunity to study the topic (in school). The proportion of student responses in either category varied largely, the average being about 50%. Student responses with missing data on opportunity to learn were omitted.

The two sets of pseudo-items were analyzed simultaneously using the one, two and three-parameter logistic model as provided by the PC-BILOG program (Mislevy & Bock, 1986), assuming a distribution of the ability parameter with a mean of 0 and variance of 1 in the total sample. The main results are summarized in Figure 1, showing the scatter plots of the location parameters from the calibration using the one-parameter model (the two and three-parameter model produce essentially the same results with regard to the location parameters). Also included are the results of the likelihood-ratio tests to test the null hypothesis that the location parameters in the two sets of pseudo-items do not differ.

Figure 1 Scatter plot of location parameters of student groups who had been taught the topic or not on three IEA subtests



On the subtest Estimation, which is supposed to be the least curriculum dependent of the three, the location parameters in both item sets do not differ significantly ($p > .90$). The same holds for the subtest Plane Figures, although the probability is much lower ($p > .20$). On the third subtest, Algebra, which like the Plane Figures test can supposed to be curriculum dependent to a large extent, the location parameters differ significantly ($p < .001$). This is also the most difficult test: p-values ranging from .12 to .55 with a mean of .31. The subtest Plane Figures is slightly less difficult (mean p-value: .43; range: .13 - .74), while the subtest Estimation is the easiest one (mean p-value: .58; range: .28 - .91).

These results are rather puzzling. The differences in outcomes cannot be explained by the degree of curriculum dependency, but they cannot be explained by differences in difficulty, either. A re-analysis of the subtests Algebra and Plane Figures in which the most difficult items were omitted yielded essentially the same results.

The findings from these exploratory analyses are more or less comparable to the results reported by Wilmot and Tuson (1983). Although probably the proportion of students not exposed to the topics tested in the Dutch assessment program will be rather small, there is reason for care. Therefore, in the calibration phase the stability of the item parameters will be checked.

References

Bock, R.D., & Mislevy, R.J. (1981). An item response curve model for matrix-sampling data: The California grade-three assessment. In: D. Carlson (Ed.), Testing in the states: Beyond accountability (New Directions for Testing and Measurement, no. 10, pp. 65-90). San Francisco: Jossey-Bass.

Bock, R.D., Mislevy, R., & Woodson, C. (1982). The next stage in educational assessment. Educational Researcher, 11 (3), 4-11, 16.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Messick, S., Beaton, A., & Lord, F. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era. Princeton, NJ: National Assessment of Educational Progress.

Mislevy, R.J. & Bock, R.D. (1986). PC-BILOG: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Pandey, T.N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In: R.K. Hambleton (Ed.), Applications of item response theory (pp. 212-229). Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R.E., & Wolfe, R.G. (1981). Latent trait theories and the assessment of educational achievement. In: D.C. Berliner (Ed.), Review of research in education 9 (pp. 377-435). Washington, DC: American Educational Research Association.

Wilmott, A.S., & Tuson, J. (1983). Exposure and performance: An investigation of test validity within the context of the APU monitoring programme. London: Assessment of Performance Unit, Department of Education and Science.